

eda部分报告

👉 下面是标准版报告笔记（照你说的标准）

📖 Step 1: Exploratory Data Analysis (EDA) — 总结笔记

1.1 Load Data

- 📦 成功读取数据，基本展示了数据头部（df.head()）和结构（df.info()）。
 - ✅ 没有问题。
-

1.2 Variable Types


- 🔍 对每一列进行了数据类型检测（object, int64, float64）。
 - 🛠️ 发现 `DATE_DIED` 是 object 类型，需要特别处理（后面生成了 `DIED` 列，解决了问题）。
 - ✅ 暂时没有问题，但后续如果新特征出现 object 类型，建模前必须转换成数值型。
-

1.3 Missing Values Detection



- 🔍 检查了缺失值（定义为97/98/99），以及 `DATE_DIED` 中的 "9999-99-99"。
 - 📊 画出了缺失值数量条形图。
 - ⚠️ 注意：这里只是统计了缺失值，填补操作并没有在这里进行，而是安排在 1.10。
-

1.4 Outlier Detection (1D Analysis)





- 📏 使用IQR方法（ $Q1 - 1.5IQR \sim Q3 + 1.5IQR$ ）检测了 `AGE` 的异常值。
- 📈 绘制了 `AGE` 的BoxPlot，图中小圈圈是异常值（outliers），合理存在。
- 🧑 发现87.5岁以上的人群属于异常范围。

-  检测完成，暂时没有进一步处理（异常值没有删除，只是识别）。
-



1.5 Variable Distributions

-  单变量分布图绘制完成（AGE直方图 / 分类变量countplot）。
 -  没有问题。
-



1.6 Correlation Analysis

-  计算了特征之间的相关系数（Correlation Matrix）。
 -  绘制了热力图（heatmap）。
 -  重点观察了 AGE、TEST_RESULT与其他特征的相关性。
 -  没有问题。
-




1.7 Important Feature Recognition

-  根据相关性分析 + 经验常识，初步选定了 O1/O2/O3 重要特征。
 -  特征选择是基于直觉和初步EDA，后续建模时可能进一步精炼。
-


1.8 Feature Means Grouped by Death (分组对比)




-  计算了死亡与未死亡组在重要特征上的均值差异。
 -  缺失值暂时是简单填补之后计算的，所以结果**偏保守**。
-

1.9 Baseline Model (Simple Decision Tree)

-  训练了一个简单的 DecisionTree（max_depth=3）作为baseline。
 -  打印了 Precision, Recall, F1-score 等指标。
 -  填补方式仍是简单众数/中位数，后续可能影响模型质量，需要优化。
-

1.10 Missing Value Imputation

-  统一处理了所有缺失值：
 - 分类变量（nunique<=20）用**众数（mode）**填补。

- 数值变量（连续型）用**中位数（median）**填补。
 -  填补方法较粗糙，在正式建模阶段（O1/O2/O3）需要优化。
 - 计划使用 KNNImputer 或 小模型预测缺失。
 -  保存了处理后的干净数据（`custom_covid19_cleaned.csv`）。
 -  处理成功。
-

总体EDA小总结

- 基本覆盖了EDA所有必做步骤：数据读取 → 缺失检查 → 异常检测 → 分布观察 → 相关性分析 → 重要特征识别。
- 当前缺点：
 - 缺失值填补比较粗糙（众数/中位数，适合EDA但不适合最终建模）。
 - 异常值（outliers）仅做了识别，暂时没有处理。
- 在后续建模阶段（Step 2~Step 4）必须针对缺失值填补和异常值处理做进一步完善。